UDK [575:61]:004.6

ADAPTABLE GENETIC ALGORITHM FOR PRIMARY PROCESSING OF MEDICAL DATA BASED ON RANDOM SEARCH

A.K.Nishanov
Professor, Information Technologies Software Department, Tashkent University of Information Technologies named after Muhammad al-Khorazmi, Uzbekistan
nishanov.a@tuit.uz

G.P.Jurayev
Independent researcher, Center for Information and Communication Technologies at the Tashkent University of Information Technologies named

N.B.Narziyev
Senior teacher, Information Technologies Software Department, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
n.b.narziyev@gmail.com

Abstract. In the article the topical software tool "Algorithmic and computer software for the primary processing of medical data" has been developed, and based on it several practical issues related to cardiology have been solved in visualized way.

The software is based on methods and algorithms for solving such issues as preprocessing of data generated by medical professionals, re-classifying study facilities and determining the importance of classes, creating table charts, and selecting a set of informative symbols that differentiate between class objects.

Keywords: Fishers Criterion, Feature Selection, Classification, Algorithms for an estimate calculation, Preprocessing of medical data, algorithm "A", flexible genetic algorithm based on random search.

I. Introduction

With the ever-increasing amount of information about objects, processes and events in various areas of human activity around the world, great attention is being paid to the development and improvement of intellectual analysis systems.

One of the topical issues of modern, information-based health care systems is the transition from large-scale character space to small-scale medical data processing, namely the development of algorithms for identifying the most important symbols, and the classification of issues using them.

The software developed by the medical experts includes data processing, pre-classification of chosen objects and determination of classification levels [1, 4, 9,10], benchmarking, and selection of informative symbols that differentiate between class objects and finding a solution. The problems and issues most commonly encountered in most of the literature reviewed [5-8, 11-13] are based on the Fisher-type criteria for finding solutions. In addition, a tool for calculating the complexity of the problem under consideration was used in the calculation algorithms. [1-13].

The interface of the software "Algorithmic and computer software for the primary processing of medical data" consists of 6 modules called "Primary Data Reading", "Algorithms for Primary processing of Medical Information", "Classification of Medical Information", adaptive genetic algorithm based on random search "," Classification based on informational character "," General information about software ".

I. “Preliminary Data Readings”. The mechanism of this module reads the initial data prepared in the Exel program, which is displayed in the tutorial program window, indicating the need for the next step.

The activities of the software are one of the most important first steps in forming the learning sample based on primary data. Therefore, it is necessary to clarify the issue.

Considering the issue. Let's assume that the learning sample formed on the basis of primary data is divided into the elective classes and given as follows:

$$K_1 = \begin{bmatrix} x_{11}^1 & x_{11}^2 & \dots & x_{11}^N \\ x_{12}^1 & x_{12}^2 & \dots & x_{12}^N \\ \vdots & \vdots & \vdots & \vdots \\ x_{1m_1}^1 & x_{1m_1}^2 & \dots & x_{1m_1}^N \end{bmatrix} \dots K_r = \begin{bmatrix} x_{r1}^1 & x_{r1}^2 & \dots & x_{r1}^N \\ x_{r2}^1 & x_{r2}^2 & \dots & x_{r2}^N \\ \vdots & \vdots & \vdots & \vdots \\ x_{rm_r}^1 & x_{rm_r}^2 & \dots & x_{rm_r}^N \end{bmatrix}.$$

This can be summarized as follows:

$$K_p = \begin{bmatrix} x_{p1}^1 & x_{p1}^2 & \dots & x_{p1}^N \\ x_{p2}^1 & x_{p2}^2 & \dots & x_{p2}^N \\ \vdots & \vdots & \vdots & \vdots \\ x_{pm_p}^1 & x_{pm_p}^2 & \dots & x_{pm_p}^N \end{bmatrix}$$

Here $p = \overline{1,r};$ as well the learning sample expressed as $K = \bigcup_{p=1}^{r} K_p$, they should be composed of non-intersecting classes, that is $K_p \cap K_q = \emptyset, (p \neq q, p = \overline{1,r};\ q = \overline{1,r};)$ terms are given.

Likewise, the components of object x_{pi} – consist of real numbers x_{pi}^{j}-, and it is read as follows: p – related to class i – patient j - symbol. Here $p = \overline{1,r};, i = \overline{1,m_p};, j = \overline{1,N};$ as well as, r is the total number of classes given, m_p- p – total number of patients in the class and N – denotes the total number of characters.

In the issues we are looking at, each class is treated as one type of disease, that is, K_1-class is “Severe angina” (number of objects - 140), K_2-class is “Acute myocardial infarction” (number of objects is 120), Class K^3 "Arithmetic form" (number of objects - 40), K^4 class is "Cardiosclerosis after infarction" (number of objects - 35). At the same time, the character space that characterizes each class (type of disease) is formed by experts in the field, consisting of 62 signs that characterize each class.

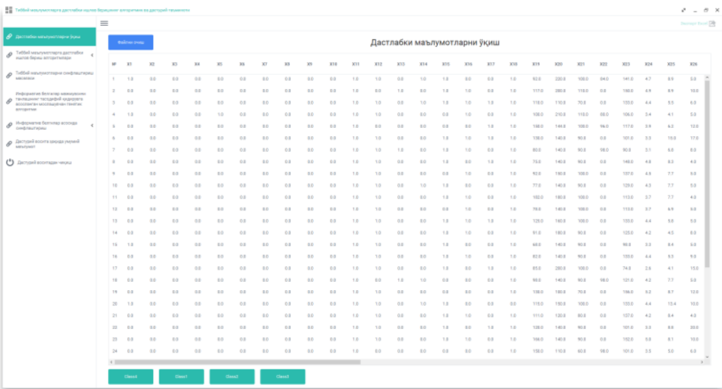


Figure *1*. – View of the "Initial Data Reading" Module window

II. "Algorithms of Medical Data Initial Processing" is implemented in two steps: 1) determining the level of suitability of the data; 2) Converting the value of the characteristic of objects to 0 or 1.

Stage 1: In the data validation phase, the characterization criteria for each K_p class are evaluated in the following order:

a). Let's make the following notes : $\bar{x}_p = (\bar{x}_p^1, \bar{x}_p^2, \ldots, \bar{x}_p^N)$ vector, X_p middle class objects, $p = \overline{1, r}$. Calculate its components by the following formula:

$$\bar{x}_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} x_{pi}^j, \; p = \overline{1,3}; \; j = \overline{1,62}; \; i = \overline{1, m_p}. \quad (1)$$

b) The distance between x_{pi} *and* $\bar{x}_p$ objects of Xp class is calculated by the following formula:

$$|x_{pi} - \bar{x}_p| = \sqrt{\sum_{j=1}^{N} (\bar{x}_p^j - x_{pi}^j)^2}, \; p = \overline{1,4}; \; j = \overline{1,62}; \; i = \overline{1, m_p}.$$

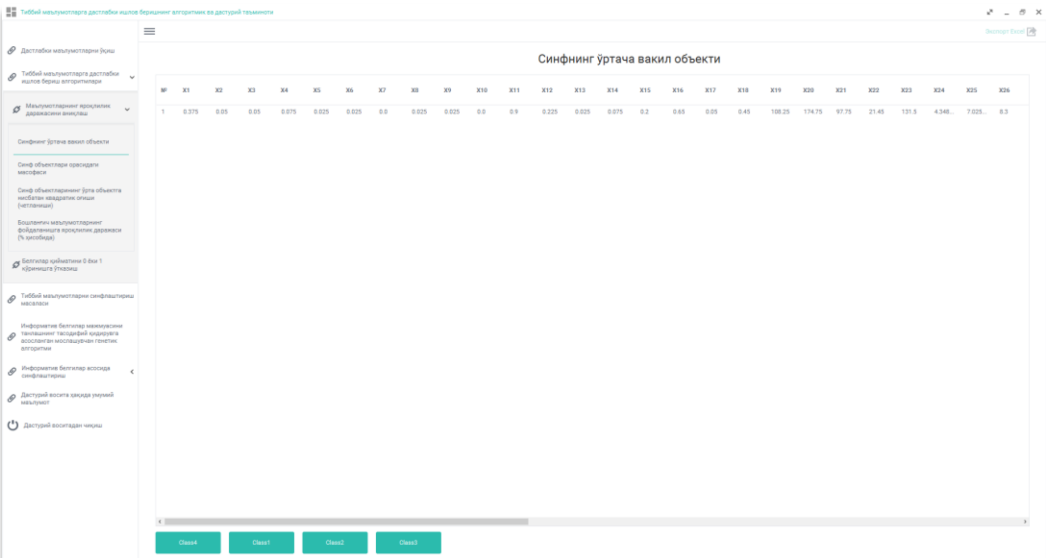
c). Quadratic deviation (deviation) of the Xp-class objects relative to the middle object $D(\bar{x}_p)$ is calculated by the following formula:

$$D(\bar{x}_p) = \sqrt{\frac{1}{\mathrm{m_p}} \sum_{j=1}^{m_p} |x_{pi} - \bar{x}_p|^2} = \sqrt{\frac{1}{m_p} \sum_{i=1}^{m_p} \sum_{j=1}^{N} (\bar{x}_p^j - x_{pi}^j)^2}, \; p = \overline{1,4}; \; j = \overline{1,62}; \; i = \overline{1, m_p}.$$

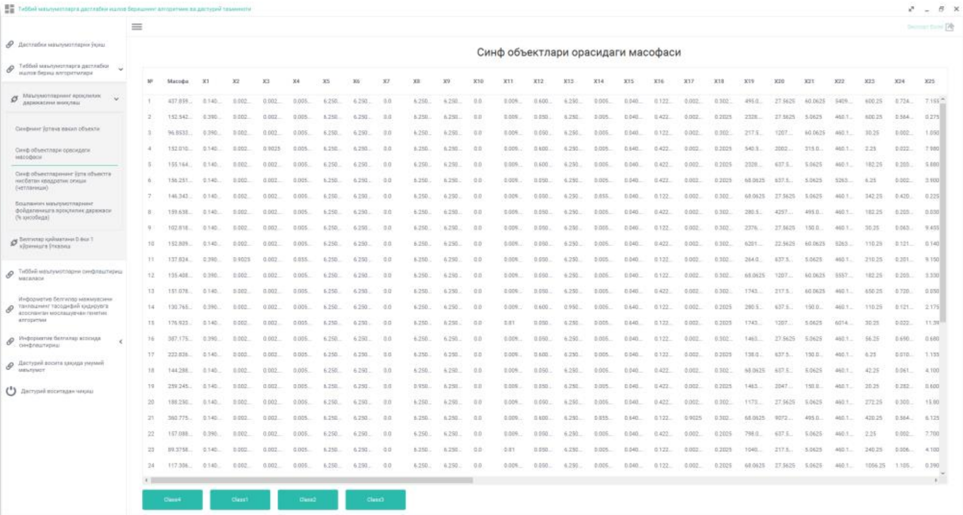
d). the following inequality is met and its performance is calculated as a percentage of class objects:

$$|x_{pi} - \bar{x}_p| \leq D(\bar{x}_p), \; p = \overline{1, r}; \; i = \overline{1, m_p}.$$

At the end of the process, the usability of each class is calculated as a percentage.

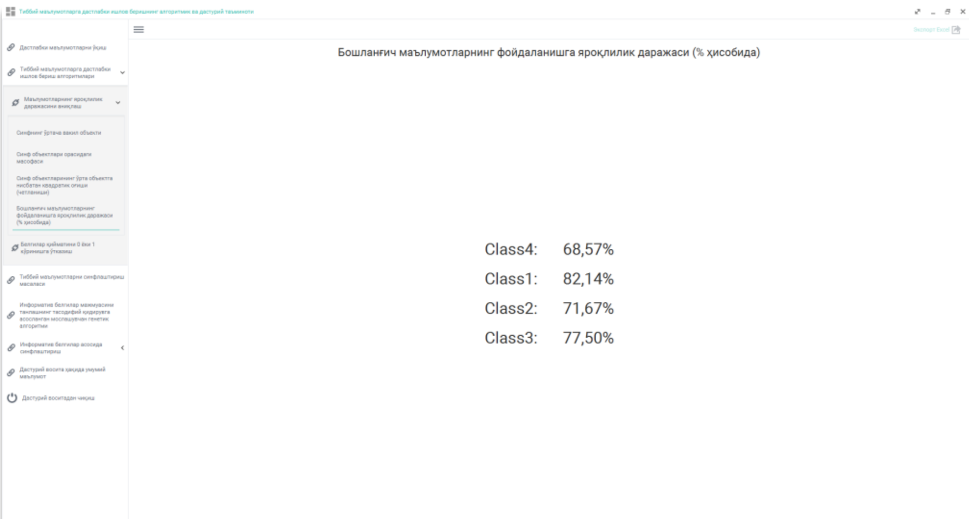


a)



b)

c)



d)

Figure 2. – View the Data Validation Dialog Box(a,b,c,d)

Stage 2: The initial data is presented in a continuous quantitative manner, during which the process of converting the character values of the class objects to values of 0 and 1 vector.

The process of converting the values of zero or one-character symbols to each of the above-mentioned K_p class objects in vector form is made by typing the following symbols in each class and all character sections:

a). $\bar{x}_p = (\overline{x}_p^1, \overline{x}_p^2, \dots, \overline{x}_p^N)$ vector, K_p class middle objects, $p = \overline{1,r}$. Calculate its components by the following formula:

$$\overline{x}_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} x_{pi}^j,\ p = \overline{1,r},\ j = \overline{1,N}, i = \overline{1,m_p};$$

b). Vectors $a_p = \left(a_p^1, a_p^2, \dots, a_p^N\right)$ ва $b_p = \left(b_p^1, b_p^2, \dots, b_p^N\right)$, Let us define the following and calculate its components by this formula:

$$a_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} (\overline{x}_p^j - x_{pi}^j)^2,\ p = \overline{1,r},\ j = \overline{1,N}.$$

$$b_{pi}^j = (\overline{x}_p^j - x_{pi}^j)^2,\ p = \overline{1,r},\ j = \overline{1,N}.$$

c). The components of the K_p elements of the learning sample are converted from the actual numeric form to the view using the following procedure.

$$x_{pi}^j = \begin{cases} 1\ equal, if\ \dfrac{b_{pi}^j}{a_p^j} \le 1, \\ equal\ to\ 0, otherwise. \end{cases}$$

At the end of this phase, the character values of the 4 class objects given are converted to 0 and 1 vector values.

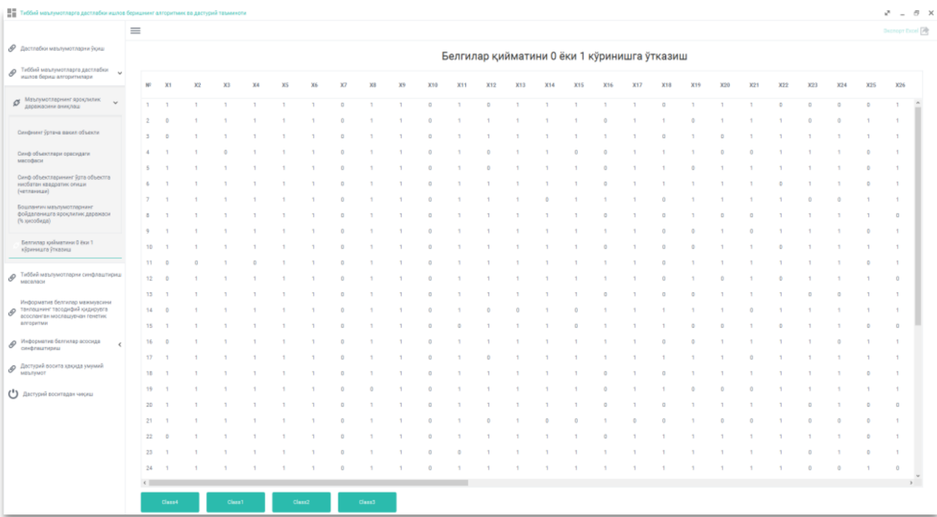


Figure *3.* – View stage conversion of character values to 0 or 1

III. Module "Classification of Medical Information" solves the problem of classification of K_p class objects, that is, whether each object in a class belongs to a class or a different one. At the same time, each object belonging to the class X_p can be compared with objects in its class and other classes by introducing $\rho_i(x_{p1,}x_{p2})$ in the informative character space:

$$\rho_i(x_{p1,}x_{p2}) = \begin{cases} 1 \quad \text{агар} \qquad (x_{p1}^i - x_{p2}^i) = 0, i = \overline{1, N}, \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

The first condition denotes the degree of similarity between the two objects, and the second condition indicates that they are different. The total of comparative evaluation is based on the following formula:

$$\Gamma_j(x_{pj,}x_{pk}) = \sum_{\text{к}=1}^{m_p} \sum_{i=1}^{N} \rho_i(x_{pj,}x_{pk}), j = \overline{1, m_p};\ k = \overline{1, m_p};\ j \neq k\,.$$

Comparative evaluation is calculated for each class, and the largest of the mean values obtained is the attribution of the object to that class.

In this section, there is a process of reclassifying classification objects to determine if a class object belongs to its own class or another class. This process is done step by step, with each step excluding objects that are not related to its class, and objects in the classes are complete until they reach their full class, ie 100%.

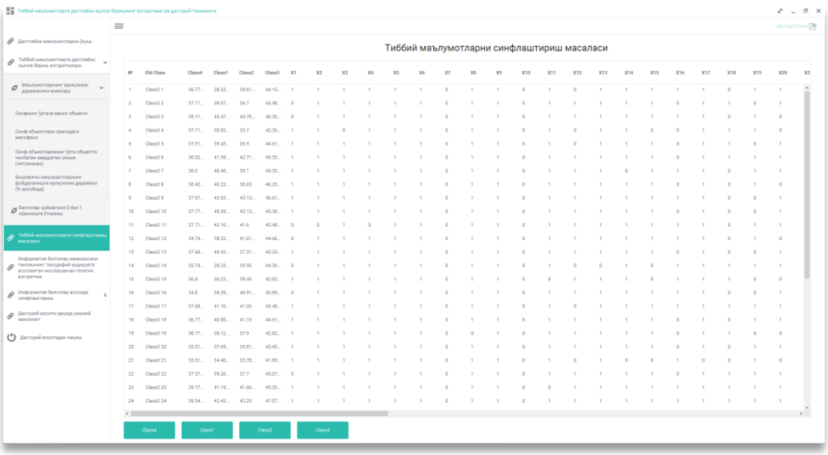


Figure *4*. – Module "Problems of Classification of Medical Information"

IV. An informative character set is selected using the reference table formulated in the module **"Random search based adaptive genetic algorithm for selecting a set of informative characters".** Let us suppose that the resultant error in classification of the sample objects based on a reference table is θ. That is, the proportion of the number of objects missing from class η to the total number of electives is set to θ.

Let the mechanism of algorithm "A" be defined as follows. Using the convergence function proposed by the authors [1, 12], the degree of importance of each sign is determined. According to it, the column will be omitted from the table with a randomly selected icon. If at the end of the process all objects in the computation find their class at least θ , the column removed from the table will not be redirected to the table otherwise the column will be returned to its original location by random selection and the process will be returned. The proposed process lasts up to ℓ . If objects have found a different class in their class (switch to another class), the randomly selected symbol will be returned. This process takes place between N characters and is separated from the remaining symbols as informative characters at the end of the process.

Hence, the essence of the work performed at this stage is to select the most useful ℓ from the set of symbols that characterize the objects being investigated, that is, the choice of an informative character tool.

The followings should be given: $\ell \ll N$ characters to be selected are the number of characters in the tool ; $p = (p^1, p^2, \ldots, p^N)$ probability that the vector components must remove one character from the character tool. Initially, $p^j = \frac{1}{N}, j = \overline{1, N}$ is taken, that is, the probability of removing an arbitrary character from an array is equal.

1-step. Algorithm "A" is used in the cross section of all marks of the given training sample. As a result, the error factor θ(N) coefficient is calculated. Typically, to determine the error rate θ(N) , the number of objects for which the error is detected is divided by the total number of objects.

2-step. Accidentally, with $p^j = \frac{1}{N}, j = \overline{1, N}$ probability, one character is selected from □ characters, and it is omitted from the array. Then the algorithm "A" is run at the intersection of symbols − 1 characters. As a result, the error coefficient □ (□ − 1) for symbols − 1 is calculated.

3-step. If $\theta(N) > \theta(N-1)$, in this case 2-step begins. This process is repeated h $(h \leq \ell)$ time. If the process is repeated, even if $h = \ell$ equality is met, the next step is to proceed. Likewise, if equality is not met, the process will continue until all the characters are considered one by one. After all the characters are reviewed, go to the next step.

4-step. Let us assume $h = \ell$, in this case with $p^j = \frac{1}{2N-h}$, $j = \overline{1, h}$; $p^j = \frac{2}{2N-h}, j = \overline{h+1, N}$; probability out of N symbols one symbol is chosen and it is excluded from the set. As in the second step, the process is reversed, except that the probability that the previously selected and non-existent characters that are to be randomly selected differ from each other. Then the algorithm "A" is run at the intersection of N-1 characters. As a result, the error coefficient θ (N-1) for N-1 is calculated.

5-step. If $\theta(N) > \theta(N-1)$, then 4th step begins. This process h $(h \leq \ell)$ is repeated one time and so on.

The proposed algorithm consists of two important parts: In the first part, genetics is a tool that identifies the generating characteristics of mutated objects, and the second part is a random search algorithm for selecting a new set.

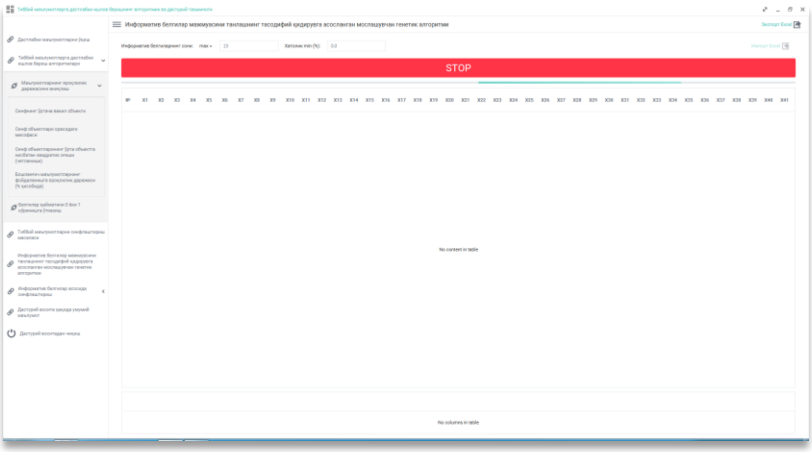


Figure 5. – Module window of "Random genetic search algorithm for selecting a set of informational symbols"

V. "Classification based on informative symbols" in the module, the classification process is carried out with the use of informative symbols, that is, in the matrix view, the columns containing informatively found symbols are retained and the remaining columns are omitted. The function of inter-object proximity in the space of the stored informative symbols is done by entering r_i (x_ (p1,) x_p2):

$$\rho_i(x_{p1,}x_{p2}) = \begin{cases} 1 \;\; if \quad (x_{p1}^i - x_{p2}^i) = 0, i = \overline{1,N}, \\ 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

The first condition denotes the degree of similarity between the two objects, and the second condition indicates that they are different. The total of comparative evaluation is based on the following formula:

$$\Gamma_j(x_{pj,}x_{pk}) = \sum_{к=1}^{m_p}\sum_{i=1}^{N}\rho_i(x_{pj,}x_{pk}), j = \overline{1,m_p};\; k = \overline{1,m_p};\; j \neq k\,.$$

Comparative evaluation is calculated for each class, and the largest of the mean values obtained is the attribution of the object to that class.

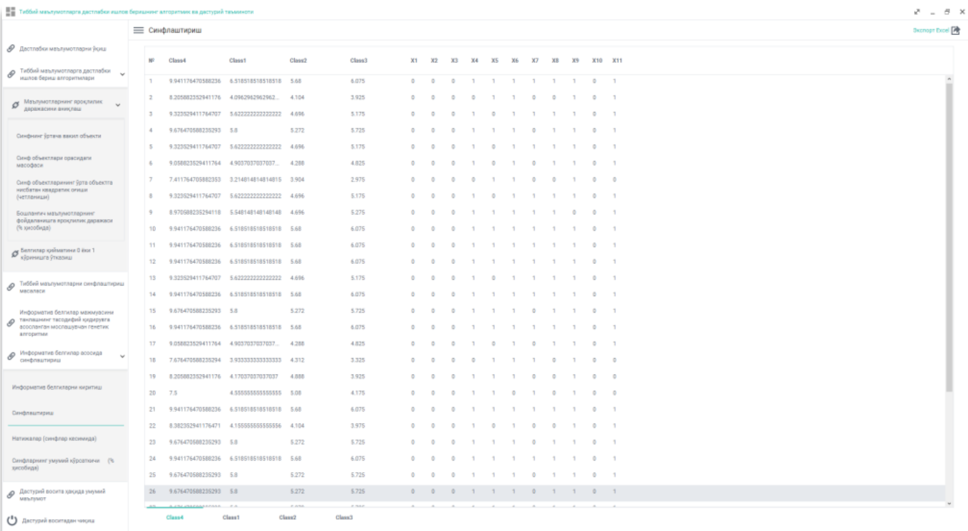


Figure *6.* – "Classification based on informational symbols" window view

VI. “General information about the software”. This module gives an overview of the software.

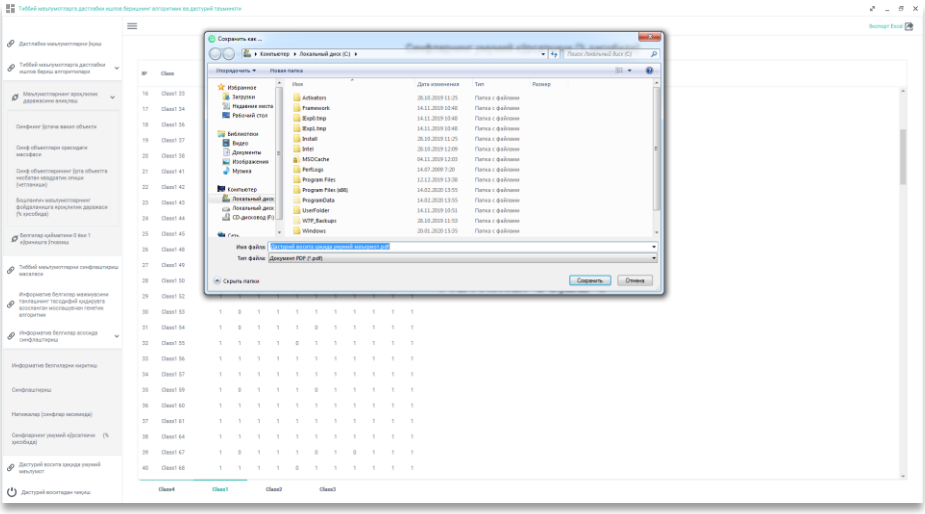


Figure *7. – «*General information about the software*»* window view

References

[1]. Zhuravlev Yu.I. Selected scientific works. – M: 'Magistr' Publishing house, 1998. - 420 p.

[2]. Yu, L., H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. – J. Mach. Learn. Res., Vol. 5, 2004, No Oct, pp. 1205-1224.

[3]. Yu, L.: Toward Integrating Feature Selection Algorithm for Classification and Clustering. IEEE Transaction on Knowledge and Data Engineering 17(4), 491-502

[4]. Yan, K., D. Zhang. Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination. – Sensors Actuators, B Chem., Vol. 212, Jun 2015, pp. 353-363.

[5]. Nishanov A.Kh,, Turakulov Kh.A., Turakhanov Kh.V. A decision rule for identification of eye pathologies // Biomedical Engineering (1999) 33(4) 178-179.

[6]. Nishanov A. Kh,, Turakulov Kh.A., Turakhanov Kh.V. A decisive rule in classifying diseases of the visual system // Meditsinskaia tekhnika (1999) (4) 16-18.

[7]. Xiang Fang, Lina Wang Feature Selection Based on Fisher Criterion and Sequential Forward Selection for Intrusion Detection // Revista de la Facultad de Ingeniería U.C.V., Vol. 32, N° 1, pp.498-503, 2017.

[8]. Linhui Sun, Sheng Fu and Fu Wang. Decision tree SVM model with Fisher feature selection for speech emotion recognition // Sun et al. EURASIP Journal on Audio, Speech, and Music Processing (2019) 2019:2 https://doi.org/10.1186/s13636-018-0145-5

[9]. Yan, X.,Tan, M., Yan,Y., Lü, M. (2012). "Research on Hidden Markov Model-Based And Neural Network-Based Intrusion Detections". Computer Applications and Software. 29(2), 294–297.

[10]. JING, X.,WANG, H., NIE, K., LUO, Z. (2012). "Feature Selection Algorithm Based on IMGA and MKSVM to Intrusion Detection". Computer Science, 39(7), 96–100.

[11]. Koushal Kumar, Jaspreet Singh Batth. Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms // International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016, 1-13.

[12]. Nishanov A.Kh., Djurayev G.P., Kasanova M.Kh. Improved algorithms for calculating evaluations in processing medical data // National Institute of Science Communication and Information Resources (NISCAIR)-India, 2019,-3158-3165.

[13]. Kamilov M., Nishanov A., Beglerbekov R. Modified stages of algorithms for computing estimates in the space of informative features // International Journal of Innovative Technology and Exploring Engineering (2019) 8(6).